AUTHOR(S): KALICIAK, L., MYRHAUG, H., GOKER, A. and SONG, D.

TITLE: Early fusion and query modification in their dual late fusion forms.

YEAR: 2015

Early Fusion and Query Modification in Their Dual Late Fusion Forms

Leszek Kaliciak, Hans Myrhaug, Ayse Goker, and Dawei Song

***Abstract*—In this paper, we prove that specific widely used models in Content-based Image Retrieval for information fusion are interchangeable. In addition, we show that even advanced, non-standard fusion strategies can be represented in dual forms. These models are often classified as representing early or late fusion strategies. We also prove that the standard Rocchio algorithm with specific similarity measurements can be represented in a late fusion form.**

***Index Terms*—Information and data fusion, early fusion, late fusion, Content-based Image Retrieval, Information Retrieval, Multimedia Retrieval, textual representation, visual representation, hybrid relevance feedback.**

I. Introduction

FUSION strategies[1] play an important role in many areas of research, including text Information Retrieval (IR), Content-based Image Retrieval (CBIR), Computer Vision, Geospatial Information Systems, Business Intelligence, Bioinformatics - to name a few. In CBIR and Computer Vision, the most widely used fusion schemes are early and late fusion strategies. They are important because they allow us to combine various notions of visual information, textual information, etc. at the representation level or system decision level.

In general, Content-based Image Retrieval is usually based on the Vector Space Model. It represents information objects as multidimensional vectors. A user query is also represented as a vector which can be an image (referred to as visual example) or text. It often contains two types of information - visual and textual. When the user submits his/her query, the similarity measurement is applied to compute the relevance scores denoting the similarities between the query and images in the data collection. The images are then ranked according to the relevance scores and the top n images are presented to the user.

Based on the experimental results, researchers have hinted at the potential interchangeability of specific fusion schemes ([14]). In this paper, we mathematically prove that this interchangeability is directly related to the interaction between early fusion operators and similarity measurements. Thus, we validate the hypotheses (interchangeability of specific fusion approaches) that stem from experimental observations and show the equivalence of particular fusion models. In addition, we also derive equivalent, dual forms of the Rocchio query modification model.

This journal paper is an extension and a follow-up of our previous papers ([9], [10]). Here, we enrich the original publication with specific non-standard early fusion strategies and show that even advanced models based on the early and late fusion strategies can be interchangeable. We also devote an entire section to proving that the standard Rocchio query modification model ([20], [1], [34]) has its late fusion form equivalent, a dual form - which would differ with respect to the similarity measure. The late fusion analogues to the Rocchio algorithm had so far been considered as separate, different techniques ([22], [12]). Section related to hybrid relevance feedback model is based on another conference publication [10].

The rest of this paper is organized as follows: Section 2 presents the background and related work on the early and late fusion schemes. Section 3 shows the relationships between various models representing different fusion strategies with examples. An interesting finding which presents dual late fusion forms of the standard Rocchio query modification model can be found in Section 4. Finally, conclusions and information on future work are provided in Section 5.

II. Background and Related Work

Different features (i.e. various visual, textual) in CBIR represent complementary yet correlated aspects of the same multimedia objects. This, in turn, presents an opportunity to utilize this complementarity by combining the feature spaces in order to improve their performance. Fusion strategies are the main tools that can be used to accomplish the aforementioned task ([5], [3]). Early fusion strategy combines the feature spaces at the representation level (fusion of representations) whereas late fusion strategy combines them at the decision level (fusion of relevance scores). Thus, for example, one can combine visual and textual features in the first round retrieval. It is also possible to combine them in the context of user relevance feedback ([10],[8],[6]). Moreover, search by multiple visual examples also requires combination of features corresponding to these visual examples. Many current state-of-the-art CBIR systems combine various visual features (often local and global) to achieve the best performance (e.g. [17]).

The most widely used early fusion technique is concatenation of visual and textual representations. In fact, some researchers implicitly assume concatenation to be synonymous with an early fusion strategy. Other recently proposed models incorporate the tensor product to combine visual and textual systems [30]. The tensor product represents a useful fusion technique as it takes into account all the combinations of different features' dimensions. It has also other applications in IR, for example, to model semantic (verb-noun pairs)

[1]In this paper, terms "fusion strategies", "fusion schemes", "fusion methods", and "fusion techniques" will be used interchangeably.

composition [31]. The main drawback of the early fusion approach is the well-known curse of dimensionality. Later in the paper we show, that the curse of dimensionality can be avoided if the equivalent late fusion form is known.

In the case of the late fusion, the most widely used method is the arithmetic mean of the scores, their sum (referred to as CombSUM in the literature), or their weighted linear combination. One of the best performing systems on the ImageCLEF2007 data collection, XRCE [17], utilizes both early - concatenation of features and late - an average of relevance scores fusion strategies for comparison purposes. Another common combination method, referred to as CombPROD in the literature, is the squared geometric mean of the relevance scores - their product. It has been argued, that the major drawback of the late fusion approaches is their inability to capture the correlation between different modalities [18]. However, in the paper we show that in some cases the late fusion can be represented in the form of an early fusion.

Early and late fusion strategies can be also considered in the context of classification, e.g. image categorization [4]. In the case of classification, late fusion is performed differently, as a weighted voting strategy from the outputs of different classiers ([21], [24]). Some fusion strategies in CBIR can be also classified as intermediate fusion [4]. They simultaneously learn individual classier and combination classier weights [33], and this process happen at various levels of learning. In this paper, however, our focus is on the similarity-based image retrieval.

Thus, in general, most literature on fusion strategies in Content-based Image Retrieval utilize either concatenation or a linear combination of relevance scores in their models (i.e. [29]). Others have used both for experimental comparison ([27], [28]) and conclude that both strategies generate similar results (slightly better performance of a late fusion) or are in favour of an early fusion strategy (i.e. [25]). All of them, however, treat these fusion strategies as separate, individual data combination approaches.

In this paper, we aim to prove that specific widely used standard and non-standard fusion models in CBIR are equivalent. All presented models are based on early and late fusion strategies, and represent counterexamples showing that these strategies should not always be considered as separate.

III. RELATIONSHIPS BETWEEN FUSION STRATEGIES

The most widely used fusion models in Content-based Image Retrieval are based on the early and late fusion schemes. We are going to show, that specific combinations of similarity measures and individual scores (late fusion) can be represented as similarities computed on pre-tensored or pre-concatenated individual representations (early fusion), and vice versa.

One of the best performing similarity measurements in information retrieval in general are: cosine similarity and metrics from the Minkowski family of distances (Euclidean, Manhattan, etc.). In particular, Euclidean distance is often utilized in visual search [11], while textual search often uses cosine similarity [35]. Moreover, late fusion is most often represented as a product of relevance scores, their sum, or their weighted linear combination ([18], [29]). The early fusion, on the other hand, is usually represented by concatenation of feature spaces ([31], [29]).

Thus, in this section, we are going to investigate the interactions between these similarity measurements and early fusion operators. We are going to reveal the relationships between concatenation and tensor product with the following similarity measurements:

- inner product[2]
- cosine similarity
- weighted cosine similarity (can be used to change the importance of different feature spaces)
- Euclidean metric

We also investigate the interactions of the aforementioned early fusion operators with a combination of Euclidean distance and cosine similarity. That is because often cosine similarity performs best in text IR (Information Retrieval) while Euclidean distance gives the best performance in CBIR. Therefore, we may want to utilize different similarity measurements for different feature spaces. Interestingly, we can combine these different similarity measurements in such a way, that this combination will correspond to the feature fusion at the representation level. Further, we explore the interactions with the Minkowski family of distances, which encompasses a wide range of various metrics and similarity measurements. The discovered relationships are supported by examples.

For the clarity of the formulas, in this section we assume that the visual and textual features were normalized. This is not a necessary assumption as analogous relationships can be found for representations that were not normalized.

Table I presents the notation used in the paper.

A. Interactions of early fusion operators (concatenation, tensor product) with the dot product

1) : We can start by making a few simple observations. Let us employ a standard inner product as the similarity measurement. Let d be a vector representation of a multimedia document. We can check that

$$\left\langle d_1^v \oplus d_1^t | d_2^v \oplus d_2^t \right\rangle = \langle d_1^v | d_2^v \rangle + \left\langle d_1^t | d_2^t \right\rangle \tag{1}$$

where $\langle \cdot | \cdot \rangle$ denotes an inner product, $\oplus$ is the direct product (concatenation of vectors) and d_i^v, d_i^t are the visual and textual image representations of the ith image, for example. We can assume that d_1^v, d_1^t denote the visual and textual query representations (query by visual example) and d_2^v, d_2^t denote the visual and textual representations of an image from the image collection. We would measure these similarities for all the images in the data collection and use the relevance scores to rank the images.

From the above equation we can see, that concatenation of vectors is equivalent to addition of measurements (scores) performed on individual feature spaces.

[2]In this paper, terms "inner product" and "dot product" will be used interchangeably.

TABLE I: Overview of the notation used in the paper.

Symbol	Meaning
d_1^v, d_1^t	Visual and textual vector representations of the query, respectively
d_2^v, d_2^t	Visual and textual vector representations of an arbitrary image from an image collection, respectively
$\oplus$	Vector concatenation (early fusion operator)
$\otimes$	Tensor product (early fusion operator)
$\langle\cdot\|\cdot\rangle$	Similarity measurement - inner product (dot product), $s_{in}(\cdot,\cdot)$
$s_c(\cdot,\cdot)$	Similarity measurement - cosine similarity
$s_e(\cdot,\cdot)$	Similarity measurement - Euclidean metric
$s_b(\cdot,\cdot)$	Similarity measurement - Bhattacharya similarity
$s_p(\cdot,\cdot)$	Similarity measurements Minkowski family of distances
$s_{in}(\cdot,\cdot)$	Similarity measurement - inner product (dot product)
$\|\|\cdot\|\|$	Vector norm
Q_d	an arbitrary document vector from the data collection
Q_m	modified query vector
Q_o	original query vector
D_j	related document vector
D_k	non-related document vector
a	original query weight
b	related documents' weights
c	non-related documents' weights
D_r	set of related documents
D_{nr}	set of non-related documents
$(.)^T$	transpose operator
A	observable
M	density matrix
$\langle A\rangle = tr(MA)$	predicted mean value of the measurement
$P = p^T p$	projector onto a subspace
Pr	probability of the projection

To clarify, concatenation ($\oplus$) of two n and m dimensional vectors produces a new $n+m$ dimensional vector, for example

$$(a,b)\oplus(c,d,e) = (a,b,c,d,e) \tag{2}$$

2) : Tensor product ($\otimes$) of two n and m dimensional vectors generates an $n\cdot m$ dimensional vector or an n by m dimensional matrix. For example

$$(a,b)\otimes(c,d,e) = (ac,ad,ae,bc,bd,be) \tag{3}$$

or

$$(a,b)\otimes(c,d,e) = \begin{pmatrix} ac & ad & ae \\ bc & bd & be \end{pmatrix} \tag{4}$$

It has been shown that the tensor product can be useful when combining the representations as it takes into account all of the combinations of vectors' dimensions and gives good discrimination in terms of similarity measurements [13]. Assuming that the systems were prepared independently, we have

$$\left\langle d_1^v\otimes d_1^t | d_2^v\otimes d_2^t\right\rangle = \langle d_1^v|d_2^v\rangle\cdot\left\langle d_1^t|d_2^t\right\rangle \tag{5}$$

where $\otimes$ denotes the tensor operator.

From the above equation it turns out that the inner product of the tensor products is a product of the measurements (scores) performed on individual feature spaces. One of the implications of this observation is that there is no need for performing the tensor operation.

B. Interactions of early fusion operators (concatenation, tensor product) with the cosine similarity

One of the best performing similarity measures in text IR is the cosine similarity (s_c)

$$s_c(d_1,d_2) = \frac{\langle d_1|d_2\rangle}{\|d_1\|\cdot\|d_2\|} \tag{6}$$

The following equations hold

$$\begin{aligned}\left\|d_1^v\otimes d_1^t\right\| = \\ =&\sqrt{\langle d_1^v\otimes d_1^t|d_1^v\otimes d_1^t\rangle} = \\ =&\sqrt{\langle d_1^v|d_1^v\rangle\cdot\langle d_1^t|d_1^t\rangle} = \\ =&\left\|d_1^v\right\|\cdot\left\|d_1^t\right\| = \\ =&1 = \left\|d_2^v\otimes d_2^t\right\|\end{aligned} \tag{7}$$

and

$$\begin{aligned}\left\|d_1^v\oplus d_1^t\right\| = \\ =&\sqrt{\langle d_1^v\oplus d_1^t|d_1^v\oplus d_1^t\rangle} = \\ =&\sqrt{\langle d_1^v|d_1^v\rangle+\langle d_1^t|d_1^t\rangle} = \\ =&\sqrt{\left\|d_1^v\right\|^2+\left\|d_1^t\right\|^2} = \\ =&\sqrt{2} = \left\|d_2^v\oplus d_2^t\right\|\end{aligned} \tag{8}$$

Therefore, we get

$$s_c\left(d_1^v\otimes d_1^t, d_2^v\otimes d_2^t\right) = s_c\left(d_1^v,d_2^v\right)\cdot s_c\left(d_1^t,d_2^t\right) \tag{9}$$

$$s_c\left(d_1^v\oplus d_1^t, d_2^v\oplus d_2^t\right) = \frac{1}{2}\left(s_c\left(d_1^v,d_2^v\right)+s_c\left(d_1^t,d_2^t\right)\right) \tag{10}$$

Here, the square root of the similarity between the tensored representations is the geometric mean of the scores computed

independently and the similarity between the concatenated representations is the arithmetic mean of individual scores.

Let us assume, that a model incorporates cosine similarity as a measurement used in combining the sub-systems (i.e. visual features or visual and textual features). Then, the concatenation or tensor operation produces the same effect as incorporation of the CombSUM or CombPROD late fusion methods, respectively.

C. Interactions of an early fusion operator (concatenation) with the weighted cosine similarity

If we utilize weighted combinations (with r_1, r_2 denoting the weights, the importance of visual and textual representations, for example), then we get[3]

$$\begin{aligned} &s_c\left(r_1 d_1^v \oplus r_2 d_1^t, r_1 d_2^v \oplus r_2 d_2^t\right) = \\ &= \frac{1}{r_1^2 + r_2^2}\left(r_1^2 s_c\left(d_1^v, d_2^v\right) + r_2^2 s_c\left(d_1^t, d_2^t\right)\right) \end{aligned} \tag{11}$$

Proof: Because

$$\begin{aligned} \left\|(r_1 d^v) \oplus \left(r_2 d^t\right)\right\| = \\ = \sqrt{\langle (r_1 d^v) \oplus (r_2 d^t) \,|\, (r_1 d^v) \oplus (r_2 d^t)\rangle} = \\ = \sqrt{\langle r_1 d^v | r_1 d^v\rangle + \langle r_2 d^t | r_2 d^t\rangle} = \\ = \sqrt{r_1^2 \langle d^v | d^v\rangle + r_2^2 \langle d^t | d^t\rangle} = \\ = \sqrt{r_1^2 \left\|d^v\right\|^2 + r_2^2 \left\|d^t\right\|^2} = \\ = \sqrt{r_1^2 + r_2^2} \end{aligned}$$

we get

$$\begin{aligned} &s_c\left(r_1 d_1^v \oplus r_2 d_1^t, r_1 d_2^v \oplus r_2 d_2^t\right) = \\ &= \frac{\langle r_1 d_1^v \oplus r_2 d_1^t | r_1 d_2^v \oplus r_2 d_2^t\rangle}{r_1^2 + r_2^2} = \\ &= \frac{\langle r_1 d_1^v | r_1 d_2^v\rangle + \langle r_2 d_1^t | r_2 d_2^t\rangle}{r_1^2 + r_2^2} = \\ &= \frac{1}{r_1^2 + r_2^2}\left(r_1^2 \frac{\langle d_1^v | d_2^v\rangle}{\left\|d_1^v\right\| \left\|d_2^v\right\|} + r_2^2 \frac{\langle d_1^t | d_2^t\rangle}{\left\|d_1^t\right\| \left\|d_2^t\right\|}\right) = \\ &= \frac{1}{r_1^2 + r_2^2}\left(r_1^2 s_c\left(d_1^v, d_2^v\right) + r_2^2 s_c\left(d_1^t, d_2^t\right)\right) \end{aligned}$$

■

D. Interactions of early fusion operators (concatenation, tensor product) with the Euclidean metric

We can also find the relationships for Euclidean metric

$$s_e\left(d_1, d_2\right) = \sqrt{\langle d_1 - d_2 | d_1 - d_2\rangle}. \tag{12}$$

Thus

$$s_e\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \sqrt{s_e^2\left(d_1^v, d_2^v\right) + s_e^2\left(d_1^t, d_2^t\right)} \tag{13}$$

[3]Similar observations can be made for other similarity measurements. Here, we only present the weighted combinations for the cosine similarity.

and

$$\begin{aligned} &s_e\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = \\ &= \sqrt{s_e^2\left(d_1^v, d_2^v\right) + s_e^2\left(d_1^t, d_2^t\right) - \frac{1}{2} s_e^2\left(d_1^v, d_2^v\right) s_e^2\left(d_1^t, d_2^t\right)} \end{aligned} \tag{14}$$

Proof: (1) From the fact that

$$s_e\left(d_1, d_2\right) = \sqrt{\left\|d_1\right\|^2 + \left\|d_2\right\|^2 - 2\langle d_1 | d_2\rangle}$$

and

$$\left\|d_1 \oplus d_2\right\| = \sqrt{2}$$

we can show that

$$\begin{aligned} &s_e\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \\ &= \sqrt{\left\|d_1^v \oplus d_1^t\right\|^2 + \left\|d_2^v \oplus d_2^t\right\|^2 - 2\langle d_1^v \oplus d_1^t | d_2^v \oplus d_2^t\rangle} = \\ &= \sqrt{4 - 2\left(\langle d_1^v | d_2^v\rangle + \langle d_1^t | d_2^t\rangle\right)} = \\ &= \sqrt{2 - 2\langle d_1^v | d_2^v\rangle + 2 - 2\langle d_1^t | d_1^t\rangle} = \\ &= \sqrt{s_e^2\left(d_1^v, d_2^v\right) + s_e^2\left(d_1^t, d_2^t\right)} \end{aligned}$$

(2) Notice that

$$\begin{aligned} &s_e^2\left(d_1^v, d_2^v\right) \cdot s_e^2\left(d_1^t, d_2^t\right) = \\ &= \left(2 - 2\langle d_1^v | d_2^v\rangle\right) \cdot \left(2 - 2\left\langle d_1^t | d_2^t\right\rangle\right) = \\ &= 2\left(2 - 2\langle d_1^v | d_2^v\rangle\right) + \\ &+ 2\left(2 - 2\left\langle d_1^t | d_2^t\right\rangle\right) - 2\left(2 - 2\langle d_1^v | d_2^v\rangle \left\langle d_1^t | d_2^t\right\rangle\right) = \\ &= 2 s_e^2\left(d_1^v, d_2^v\right) + 2 s_e^2\left(d_1^t, d_2^t\right) - 2 s_e^2\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) \end{aligned}$$

■

E. Interactions of early fusion operators (concatenation, tensor product) with the Bhattacharya similarity

Similarly, for the Bhattacharya similarity

$$s_b\left(d_1, d_2\right) = -ln\left(\sum_i \sqrt{\left(d_1\right)_i \cdot \left(d_2\right)_i}\right) \tag{15}$$

we get

$$s_b\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = s_b\left(d_1^v, d_2^v\right) + s_b\left(d_1^t, d_2^t\right) \tag{16}$$

and

$$\begin{aligned} &s_b\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \\ &= -ln\left(e^{-s_b\left(d_1^v, d_2^v\right)} + e^{-s_b\left(d_1^t, d_2^t\right)}\right) \end{aligned} \tag{17}$$

Proof: Let us denote

$$\sqrt{d} = \left(\sqrt{d_1}, \sqrt{d_2}, \ldots, \sqrt{d_n}\right)$$

Then

$$\begin{aligned}
&s_b\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = \\
&= -ln\left(\sum_k \sqrt{\left(d_1^v \otimes d_1^t\right)_k \cdot \left(d_2^v \otimes d_2^t\right)_k}\right) = \\
&= -ln\left(\left\langle \sqrt{d_1^v \otimes d_1^t} | \sqrt{d_2^v \otimes d_2^t} \right\rangle\right) = \\
&= -ln\left(\left\langle \sqrt{d_1^v} \otimes \sqrt{d_1^t} | \sqrt{d_2^v} \otimes \sqrt{d_2^t} \right\rangle\right) = \\
&= -ln\left(\left\langle \sqrt{d_1^v} | \sqrt{d_2^v} \right\rangle \cdot \left\langle \sqrt{d_1^t} | \sqrt{d_2^t} \right\rangle\right) = \\
&= -\left(ln\left\langle \sqrt{d_1^v} | \sqrt{d_2^v} \right\rangle + ln\left\langle \sqrt{d_1^t} | \sqrt{d_2^t} \right\rangle\right) = \\
&= -\left(ln\sum_i \sqrt{\left(d_1^v\right)_i \cdot \left(d_2^v\right)_i} + ln\sum_j \sqrt{\left(d_1^t\right)_j \cdot \left(d_2^t\right)_j}\right) = \\
&= s_b\left(d_1^v, d_2^v\right) + s_b\left(d_1^t, d_2^t\right)
\end{aligned}$$

For the concatenation, we have

$$\begin{aligned}
&s_b\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \\
&= -ln\left(\sum_k \sqrt{\left(d_1^v \oplus d_1^t\right)_k \cdot \left(d_2^v \oplus d_2^t\right)_k}\right) = \\
&= -ln\left(\left\langle \sqrt{d_1^v \oplus d_1^t} | \sqrt{d_2^v \oplus d_2^t} \right\rangle\right) = \\
&= -ln\left(\left\langle \sqrt{d_1^v} \oplus \sqrt{d_1^t} | \sqrt{d_2^v} \oplus \sqrt{d_2^t} \right\rangle\right) = \\
&= -ln\left(\left\langle \sqrt{d_1^v} | \sqrt{d_2^v} \right\rangle + \left\langle \sqrt{d_1^t} | \sqrt{d_2^t} \right\rangle\right) = \\
&= -ln\left(e^{ln\left\langle \sqrt{d_1^v} | \sqrt{d_2^v} \right\rangle} + e^{ln\left\langle \sqrt{d_1^t} | \sqrt{d_2^t} \right\rangle}\right) = \\
&= -ln\left(e^{-s_b\left(d_1^v, d_2^v\right)} + e^{-s_b\left(d_1^t, d_2^t\right)}\right)
\end{aligned}$$

■

F. Interactions of early fusion operators (concatenation, tensor product) with the Euclidean Metric. Interpretation of non-linear combinations of cosine similarity and Euclidean distance

Sometimes it might be beneficial to utilize different similarity measures for different feature spaces [7] (i.e. Euclidean metric for visual features and cosine similarity for textual space). Interestingly, we can fuse the scores in such a way, that their combination would correspond to (for example) measuring the Euclidean distance between the concatenated or tensored representations

$$\begin{aligned}
&s_e\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \\
&= \sqrt{s_e^2\left(d_1^v, d_2^v\right) - 2s_c\left(d_1^t, d_2^t\right) + 2} \qquad (18) \\
&s_e\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) = \\
&= \sqrt{s_e^2\left(d_1^v, d_2^v\right) s_c\left(d_1^t, d_2^t\right) - 2s_c\left(d_1^t, d_2^t\right) + 2} \qquad (19)
\end{aligned}$$

Proof: Stems from the fact that

$$\begin{aligned}
s_e^2\left(d_1^t, d_2^t\right) = \\
&= 2 - 2\left\langle d_1^t | d_2^t \right\rangle = \\
&= 2 - 2\frac{\left\langle d_1^t | d_2^t \right\rangle}{\left\|d_1^t\right\| \left\|d_2^t\right\|} = \\
&= 2 - 2s_c\left(d_1^t, d_2^t\right)
\end{aligned}$$

and (5),(13).

■

G. Interactions of early fusion operators (concatenation, tensor product) with the Minkowski Family of Distances

Minkowski family of distances include widely utilized Manhattan and Euclidean metrics. Manhattan distance, for example, was utilized in [15] to query the CBIR system by multiple visual examples. In this aforementioned paper, individual scores corresponding to visual examples where aggregated. It is interesting to know, that if one concatenated the representations corresponding to visual examples and utilized Manhattan metric, then the influence of these fusion methods on the retrieval performance would be exactly the same.

Minkowski family of distances is represented by the formula

$$s_p\left(d_1, d_2\right) = \left(\sum_{i=1}^{n} \left|d_1^i - d_2^i\right|^p\right)^{\frac{1}{p}} \qquad (20)$$

where $p \in \mathbb{N}$.

For the fractional values of $p \in (0, 1)$, the formula is not a metric in the mathematical sense. However, it has been shown [16], that the similarity measure with fractional values of p works well in CBIR.

We are going to show, that

$$s_p\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \left(s_p^p\left(d_1^v, d_2^v\right) + s_p^p\left(d_1^t, d_2^t\right)\right)^{\frac{1}{p}} \qquad (21)$$

Proof:

$$\begin{aligned}
&s_p\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) = \\
&= \left\|d_1^v \oplus d_1^t - d_2^v \oplus d_2^t\right\|_p = \\
&= \left\|\left(d_1^v - d_2^v\right) \oplus \left(d_1^t - d_2^t\right)\right\|_p = \\
&= \left(\left\|d_1^v - d_2^v\right\|_p^p + \left\|d_1^t - d_2^t\right\|_p^p\right)^{\frac{1}{p}} = \\
&= \left(s_p^p\left(d_1^v, d_2^v\right) + s_p^p\left(d_1^t, d_2^t\right)\right)^{\frac{1}{p}}
\end{aligned}$$

where

$$\left\|d\right\|_p = \left(d_1^p + d_2^p + \ldots d_n^p\right)^{\frac{1}{p}}$$

Here, the representations do not have to be normalized.

■

Hence, in these cases the early and late fusion approaches are interchangeable. The fusion of representations is then, in fact, the fusion of similarities computed independently on visual and textual feature spaces. This is, in our opinion, an interesting finding.

H. Advanced Early Fusion and Interchangeability

The following section is based on and contains excerpts from [10].

Even advanced, non-standard early fusion can in some cases be represented as a late fusion. The hybrid CBIR relevance model introduced in [10] can be considered as a dual form fusion. The model is based on the tensor product of co-occurrence matrices representing visual and textual subspaces of queries and feedback images. It was proven that this advanced measurement performed on the combined representations is equivalent to the non-trivial combination of measurements performed on individual feature spaces. Knowledge of this interchangeability makes the models easy to implement and significantly faster (computations performed on individual feature spaces).

Modern retrieval systems allow the users to interact with the system in order to narrow down and refine the search ([18], [10]). This interaction takes the form of implicit or explicit feedback. The representations of the images in the feedback set are often aggregated or concatenated (or co-occurrence matrices may be aggregated to represent i.e. probability distribution matrix). The information extracted from the feedback set is utilized to expand the query or re-rank the top images returned in the first round of the retrieval.

The proposed hybrid relevance feedback model was inspired by the measurement used in quantum mechanics, which is based on an expectation value, predicted mean value of the measurement

$$\langle A \rangle = tr\left(\rho A\right) \tag{22}$$

where tr denotes the trace operator, ρ represents a density matrix of the system and A is an observable. We can also represent an observable A as a density matrix (corresponding to the query or an image in the collection). For more information on the analogies between quantum mechanics and information retrieval the curious reader is referred to [23].

We are going to use the tensor operator $\otimes$ to combine the density matrices corresponding to visual and textual feature spaces. In quantum mechanics, the tensor product of density matrices of different systems represents a density matrix of the combined system (see [32]).

Thus, the proposed measurement is represented by

$$tr\left(\left(M_1 \otimes M_2\right) \cdot \left(\left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right)\right)\right) \tag{23}$$

where M_1, M_2 represent density matrices (co-occurrence matrices) of the query and images in the feedback set corresponding to visual and textual spaces respectively, a and b denote row vectors representing visual and textual information for an image from the data collection[4], and T is a transpose operation on matrices. We would perform this measurement on all the images in the collection, thus re-scoring the data collection based on the user feedback.

Assuming that the systems were prepared independently (otherwise we would have to try to model a concept analogous to entanglement [2]), we get

$$\begin{aligned} &tr\left(\left(M_1 \otimes M_2\right) \cdot \left(\left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right)\right)\right) = \\ =&tr\left(\left(M_1 \cdot \left(a^T \cdot a\right)\right) \otimes \left(M_2 \cdot \left(b^T \cdot b\right)\right)\right) = \\ =&tr\left(M_1 \cdot \left(a^T \cdot a\right)\right) \cdot tr\left(M_2 \cdot \left(b^T \cdot b\right)\right) = \\ =&\left\langle M_1 | a^T \cdot a \right\rangle \cdot \left\langle M_2 | b^T \cdot b \right\rangle \end{aligned} \tag{24}$$

where $\langle \cdot | \cdot \rangle$ denotes an inner product operating on a vector space.

Let q_v, q_t denote the visual and textual representations of the query, c^i, d^i denote visual and textual representations of the images in the feedback set, r_1, r_2 denote the weighting factors (constant, importance of query and feedback density matrices respectively), and n denote the number of images in the feedback set. Then, we define M_1 and M_2 as weighted combinations of co-occurrence matrices (a subspace generated by the query vector and vectors from the feedback set)[5]. Here, D_q^v, D_q^t, D_f^v, and D_f^t represent co-occurrence matrices of query and feedback images corresponding to visual and textual features respectively.

$$\begin{aligned} M_1 &= r_1 \cdot D_q^v + \frac{r_2}{n} \cdot D_f^v = \\ &= r_1 \cdot q_v^T \cdot q_v + \sum_i \left(\frac{r_2}{n} \cdot \left(c^i\right)^T \cdot c^i\right) \end{aligned} \tag{25}$$

and

$$\begin{aligned} M_2 &= r_1 \cdot D_q^t + \frac{r_2}{n} \cdot D_f^t = \\ &= r_1 \cdot q_t^T \cdot q_t + \sum_i \left(\frac{r_2}{n} \cdot \left(d^i\right)^T \cdot d^i\right) \end{aligned} \tag{26}$$

The common way of co-occurrence matrix generation is to multiply the term-document matrix by its transpose (rows of the matrix represent the documents $d_1, \ldots d_m$), that is $D = M^T \cdot M$. Notice, that this is equivalent to $D = \sum_{i=1}^{n} d_i^T \cdot d_i$.

This observation, due to the properties of an inner product, will allow us to further simplify our model

$$\begin{aligned} &\left\langle M_1 \otimes M_2 | \left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right) \right\rangle = \\ &= \left\langle M_1 | a^T \cdot a \right\rangle \cdot \left\langle M_2 | b^T \cdot b \right\rangle = \\ &= \left\langle r_1 \cdot q_v^T \cdot q_v + \sum_i \left(\frac{r_2}{n} \cdot \left(c^i\right)^T \cdot c^i\right) | a^T \cdot a \right\rangle \cdot \\ &= \left\langle r_1 \cdot q_t^T \cdot q_t + \sum_i \left(\frac{r_2}{n} \cdot \left(d^i\right)^T \cdot d^i\right) | b^T \cdot b \right\rangle = \\ &= \left(\left\langle r_1 \cdot q_v^T \cdot q_v | a^T \cdot a \right\rangle + \sum_i \frac{r_2}{n} \left\langle \left(c^i\right)^T \cdot c^i | a^T \cdot a \right\rangle\right) \cdot \\ &= \left(\left\langle r_1 \cdot q_t^T \cdot q_t | b^T \cdot b \right\rangle + \sum_i \frac{r_2}{n} \left\langle \left(d^i\right)^T \cdot d^i | b^T \cdot b \right\rangle\right) = \\ &= \left(r_1 \cdot \left\langle q_v | a \right\rangle^2 + \frac{r_2}{n} \cdot \sum_i \left\langle c^i | a \right\rangle^2\right) \cdot \\ &= \left(r_1 \cdot \left\langle q_t | b \right\rangle^2 + \frac{r_2}{n} \cdot \sum_i \left\langle d^i | b \right\rangle^2\right) \end{aligned} \tag{27}$$

[4]For the clarity of formulas $a = d_2^v$, $b = d_2^t$.

[5]Co-occurrence matrices are quite often utilized in the Information Retrieval (IR) field. Because they are Hermitian and positive-definite, they can be thought of as density matrices (probability distribution).

Notice that the model breaks down into the weighted combinations of individual measurements. The squares of the inner products come from the correlation matrices and can play an important role in the measurement. Later in the paper, we are going to justify this claim.

We can consider a variation of the aforementioned model, where just like in the original one $M_1 = r_1 \cdot D_q^v + \frac{r_2}{n} \cdot D_f^v$ and $M_2 = r_1 \cdot D_q^t + \frac{r_2}{n} \cdot D_f^t$. We can decompose (eigenvalue decomposition) the density matrices M_1, M_2 to estimate the bases[6] (p_i^v, p_j^t) of the subspaces generated by the query and the images in the feedback set. Now, let us consider the measurement

$$\left\langle P_1 \otimes P_2 \right| \left(a^T a\right) \otimes \left(b^T b\right) \rangle \tag{28}$$

where P_1, P_2 are the projectors onto visual and textual subspaces generated by query and the images in the feedback set ($\sum_i \left(p_i^v\right)^T p_i^v$, $\sum_j \left(p_j^t\right)^T p_j^t$), and a, b are the visual and textual representations of an image from the data set. Because the tensor product of the projectors corresponding to visual and textual Hilbert spaces (H_1, H_2) is a projector onto the tensored Hilbert space[7] ($H_1 \otimes H_2$), the model can be interpreted as probability of relevance context, the probability that vector $a \otimes b$ was generated within the subspace (representing the relevance context) generated by $M_1 \otimes M_2$. Hence

$$\begin{aligned}
&\left\langle P_1 \otimes P_2 \right| \left(a^T a\right) \otimes \left(b^T b\right) \rangle = \\
&= \left\langle P_1 | a^T a \right\rangle \cdot \left\langle P_2 | b^T b \right\rangle = \\
&= \left\langle \sum_i \left(p_i^v\right)^T p_i^v | a^T a \right\rangle \cdot \left\langle \sum_j \left(p_j^t\right)^T p_j^t | b^T b \right\rangle = \\
&= \sum_i \left\langle p_i^v | a \right\rangle^2 \cdot \sum_j \left\langle p_j^t | b \right\rangle^2 = \\
&= \sum_i Pr_i^v \cdot \sum_j Pr_j^t = \\
&= \left\| \left(\left\langle p_1^v | a \right\rangle, \ldots, \left\langle p_n^v | a \right\rangle\right) \otimes \left(\left\langle p_1^t | b \right\rangle, \ldots, \left\langle p_n^t | b \right\rangle\right) \right\|^2
\end{aligned} \tag{29}$$

where Pr denotes the projection probability and $\|\cdot\|$ represents vector norm.

We can see that this measurement is equivalent to the weighted combinations of all the probabilities of projections for all the images involved. In quantum mechanics, the square of the absolute value of the inner product between the initial state and the eigenstate is the probability of the system collapsing to this eigenstate. In our case, the square of the absolute value of the inner product can be interpreted as a particular contextual factor influencing the measurement.

[6]It has been highlighted [19] that the orthogonal decomposition may not be the best option for visual spaces because the receptive fields that result from this process are not localized, and the vast majority do not at all resemble any known cortical receptive fields. Thus, in the case of visual spaces, we may want to utilize decomposition methods that produce non-orthogonal basis vectors.

[7]A Hilbert space is a vector space with an inner product operation on elements of the vector space. It is a generalization of the notion of a Euclidean space. Hence, Hilbert spaces allow us to utilize a wider variety of mathematical tools to model various phenomena in IR, for example. This generalization can also often encompass many different models operating in Euclidean space, thus unifying various approaches.

IV. Query Modification and Late Fusion

Query reformulation techniques are often used in multimedia retrieval to narrow down the search based on the user feedback. We are going to show, that the Rocchio query modification algorithm [26] can be represented as a late fusion, a combination of a number of individual relevance scores. This interesting finding shows that the same effect can be achieved by either modifying the query or combining individual relevance scores.

The late fusion analogues to the Rocchio algorithm have been considered as separate, different techniques ([22], [12]). We show that one of the standard query modification algorithms, the Rocchio model, also has its dual late fusion form representations.

Rocchio algorithm modifies the query so that it moves closer to the centroid of relevant documents and further away from the centroid of irrelevant ones

$$\begin{aligned}
Q_m = (a \cdot Q_o) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{D_j \in D_r} D_j\right) - \\
\left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{D_k \in D_{nr}} D_k\right)
\end{aligned} \tag{30}$$

where
Q_m - modified query vector
Q_o - original query vector
D_j - related document vector
D_k - non-related document vector
a - original query weight
b - related documents' weights
c - non-related documents' weights
D_r - set of related documents
D_{nr} - set of non-related documents

We will show, that the modification of the query can be interpreted as a weighted combination of the measurements (scores, similarities) between a query and a document from the data collection and between a query and each document from the feedback set. In this section, for the clarity of the formulas, we assume that all vectors were normalized to unit vectors and Q_d denotes an arbitrary document vector from the data collection.

A. Inner Product

After modifying the query, we need to re-compute the scores. Thus, we would get

$$\begin{aligned}
\langle Q_m | Q_d \rangle = \\
= a \langle Q_o | Q_d \rangle + \frac{b}{|D_r|} \sum_{D_j \in D_r} \langle D_j | Q_d \rangle - \\
\frac{c}{|D_{nr}|} \sum_{D_k \in D_{nr}} \langle D_k | Q_d \rangle
\end{aligned} \tag{31}$$

Proof:

$$\begin{aligned}
&\langle Q_m|Q_d\rangle = \\
&= \left\langle aQ_o + b\frac{1}{|D_r|}\sum_{D_j\in D_r} D_j - c\frac{1}{|D_{nr}|}\sum_{D_k\in D_{nr}} D_k | Q_d \right\rangle = \\
&= \langle aQ_o|Q_d\rangle + \\
&+ \left\langle \frac{b}{|D_r|}\sum_{D_j\in D_r} D_j|Q_d \right\rangle - \left\langle \frac{c}{|D_{nr}|}\sum_{D_k\in D_{nr}} D_k|Q_d \right\rangle = \\
&= a\langle Q_o|Q_d\rangle + \frac{b}{|D_r|}\sum_{D_j\in D_r}\langle D_j|Q_d\rangle - \\
&\frac{c}{|D_{nr}|}\sum_{D_k\in D_{nr}}\langle D_k|Q_d\rangle
\end{aligned}$$

■

Hence, the query modification with the inner product as a similarity measurement can be represented in a specific late fusion form.

B. Cosine Similarity

For the cosine similarity, we get

$$\begin{aligned}
&s_c(Q_m, Q_d) = \frac{1}{\|Q_m\|}\Bigg(as_c(Q_o, Q_d) + \\
&+ \frac{b}{|D_r|}\sum_j s_c(D_j, Q_d) - \frac{c}{|D_{nr}|}\sum_k s_c(D_k, Q_d)\Bigg)
\end{aligned} \tag{32}$$

$$\begin{aligned}
&\|Q_m\|^2 = a^2 + c^2 + \\
&+ \frac{2ab}{|D_r|}\sum_j s_c(Q_o, D_j) - \frac{2ac}{|D_{nr}|}\sum_k s_c(Q_o, D_k) - \\
&\frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k s_c(D_j, D_k)
\end{aligned} \tag{33}$$

Proof:

$$\begin{aligned}
&s_c(Q_m, Q_d) = \frac{\langle Q_m|Q_d\rangle}{\|Q_m\|\cdot\|Q_d\|} = \\
&= \frac{1}{\|Q_m\|}\Bigg(as_c(Q_o, Q_d) + \\
&+ \frac{b}{|D_r|}\sum_j s_c(D_j, Q_d) - \frac{c}{|D_{nr}|}\sum_k s_c(D_k, Q_d)\Bigg)
\end{aligned}$$

where

$$\begin{aligned}
&\|Q_m\|^2 = \langle Q_m|Q_m\rangle = \\
&= \left\langle aQ_o + \frac{b}{|D_r|}\sum_j D_j - \frac{c}{|D_{nr}|}\sum_k D_k | aQ_o + \right. \\
&\left. + \frac{b}{|D_r|}\sum_j D_j - \frac{c}{|D_{nr}|}\sum_k D_k \right\rangle = \\
&= a^2\langle Q_o|Q_o\rangle + \frac{2ab}{|D_r|}\sum_j\langle Q_o|D_j\rangle - \frac{2ac}{|D_{nr}|}\sum_k\langle Q_o|D_k\rangle - \\
&\frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k\langle D_j|D_k\rangle + \frac{c^2}{|D_{nr}|^2}\sum_k\sum_k\langle D_k|D_k\rangle = \\
&= a^2 + c^2 + \frac{2ab}{|D_r|}\sum_j s_c(Q_o, D_j) - \frac{2ac}{|D_{nr}|}\sum_k s_c(Q_o, D_k) - \\
&\frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k s_c(D_j, D_k)
\end{aligned}$$

■

Hence, the query modification with the cosine similarity as a similarity measurement can be represented in a specific late fusion form.

C. Euclidean Distance

For the Euclidean distance

$$\begin{aligned}
&s_e^2(Q_m, Q_d) = \\
&= a^2 + c^2 + 2ab - 2ac - 2bc - 2a - 2b - 2c + 1 - \\
&\frac{ab}{|D_r|}\sum_j s_e(Q_o, D_j) + \frac{ac}{|D_{nr}|}\sum_k s_e(Q_o, D_k) + \\
&+ \frac{bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k s_e(D_j, D_k) + \\
&+ as_e(Q_o, Q_d) + \\
&+ \frac{b}{|D_r|}\sum_j s_e(D_j, Q_d) + \frac{c}{|D_{nr}|}\sum_k s_e(D_k, Q_d)
\end{aligned} \tag{34}$$

Proof:

Based on the previous observation (for cosine similarity),

we get

$$
\begin{aligned}
&s_e^2\left(Q_m, Q_d\right) = \\
=&a^2 + c^2 + \frac{2ab}{|D_r|}\sum_j \langle Q_o|D_j\rangle - \\
&\frac{2ac}{|D_{nr}|}\sum_k \langle Q_o|D_k\rangle - \frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k \langle D_j|D_k\rangle + \\
+&1 - 2a\langle Q_o|Q_d\rangle - \frac{2b}{|D_r|}\sum_j \langle D_j|Q_d\rangle - \frac{2c}{|D_{nr}|}\sum_k \langle D_k|Q_d\rangle = \\
=&a^2 + c^2 + 1 + \frac{2ab}{|D_r|}|D_r| - \frac{2ab}{|D_r|}|D_r| + \frac{2ab}{|D_r|}\sum_j \langle Q_o|D_j\rangle + \\
+&\frac{2ac}{|D_{nr}|}|D_{nr}| - \frac{2ac}{|D_{nr}|}|D_{nr}| - \frac{2ac}{|D_{nr}|}\sum_k \langle Q_o|D_k\rangle + \\
+&\frac{2bc}{|D_r|\cdot|D_{nr}|}|D_r|\cdot|D_{nr}| - \\
&\frac{2bc}{|D_r|\cdot|D_{nr}|}|D_r|\cdot|D_{nr}| - \frac{2bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k \langle D_j|D_k\rangle + \\
+&2a - 2a - 2a\langle Q_o|Q_d\rangle + \\
+&\frac{2b}{|D_r|}|D_r| - \frac{2b}{|D_r|}|D_r| - \frac{2b}{|D_r|}\sum_j \langle D_j|Q_d\rangle + \\
+&\frac{2c}{|D_{nr}|}|D_{nr}| - \frac{2c}{|D_{nr}|}|D_{nr}| - \frac{2c}{|D_{nr}|}\sum_k \langle D_k|Q_d\rangle = \\
=&a^2 + c^2 + 2ab - 2ac - 2bc - 2a - 2b - 2c + 1 - \\
&\frac{ab}{|D_r|}\sum_j s_e\left(Q_o, D_j\right) + \frac{ac}{|D_{nr}|}\sum_k s_e\left(Q_o, D_k\right) + \\
+&\frac{bc}{|D_r|\cdot|D_{nr}|}\sum_j\sum_k s_e\left(D_j, D_k\right) + \\
+&a s_e\left(Q_o, Q_d\right) + \frac{b}{|D_r|}\sum_j s_e\left(D_j, Q_d\right) + \frac{c}{|D_{nr}|}\sum_k s_e\left(D_k, Q_d\right)
\end{aligned}
$$

■

Hence, the query modification with the Euclidean distance as a similarity measurement can be represented in a specific late fusion form.

D. Hybrid Relevance Feedback and Rocchio Algorithm

We can also tensor or concatenate the modified query vectors in order to generate hybrid models. Then (v, t indexes denote visual and textual representations respectively)

$$
\begin{aligned}
&\left\langle Q_m^v \otimes Q_m^t | Q_d^v \otimes Q_d^t\right\rangle = \\
=&\langle Q_m^v|Q_d^v\rangle\left\langle Q_m^t|Q_d^t\right\rangle = \\
=&(a\langle Q_o^v|Q_d^v\rangle + \frac{b}{|D_r|}\sum_{D_j\in D_r}\left\langle D_j^v|Q_d^v\right\rangle - \\
&\frac{c}{|D_{nr}|}\sum_{D_k\in D_{nr}}\langle D_k^v|Q_d^v\rangle)\cdot \\
&(a\left\langle Q_o^t|Q_d^t\right\rangle + \frac{b}{|D_r|}\sum_{D_j\in D_r}\left\langle D_j^t|Q_d^t\right\rangle - \\
&\frac{c}{|D_{nr}|}\sum_{D_k\in D_{nr}}\left\langle D_k^t|Q_d^t\right\rangle)
\end{aligned}
\tag{35}
$$

and for concatenation

$$
\begin{aligned}
&\left\langle Q_m^v \oplus Q_m^t | Q_d^v \oplus Q_d^t\right\rangle = \\
=&\langle Q_m^v|Q_d^v\rangle + \left\langle Q_m^t|Q_d^t\right\rangle = \\
=&a\left(\langle Q_o^v|Q_d^v\rangle + \left\langle Q_o^t|Q_d^t\right\rangle\right) + \\
+&\frac{b}{|D_r|}\sum_{D_j\in D_r}\left(\left\langle D_j^v|Q_d^v\right\rangle + \left\langle D_j^t|Q_d^t\right\rangle\right) - \\
&\frac{c}{|D_{nr}|}\sum_{D_k\in D_{nr}}\left(\left\langle D_k^v|Q_d^v\right\rangle + \left\langle D_k^t|Q_d^t\right\rangle\right)
\end{aligned}
\tag{36}
$$

We can use other similarity measures

$$
s_c\left(Q_m^v \otimes Q_m^t, Q_d^v \otimes Q_d^t\right) = s_c\left(Q_m^v, Q_d^v\right)\cdot s_c\left(Q_m^t, Q_d^t\right) \tag{37}
$$

$$
s_c\left(Q_m^v \oplus Q_m^t, Q_d^v \oplus Q_d^t\right) = \frac{1}{2}\left(s_c\left(Q_m^v, Q_d^v\right) + s_c\left(Q_m^t, Q_d^t\right)\right) \tag{38}
$$

$$
\begin{aligned}
&s_e\left(Q_m^v \otimes Q_m^t, Q_d^v \otimes Q_d^t\right) = \\
=&\sqrt{s_e^2\left(Q_m^v, Q_d^v\right) + s_e^2\left(Q_m^t, Q_d^t\right) - \frac{1}{2}s_e^2\left(Q_m^v, Q_d^v\right)s_e^2\left(Q_m^t, Q_d^t\right)}
\end{aligned}
\tag{39}
$$

$$
s_e\left(Q_m^v \oplus Q_m^t, Q_d^v \oplus Q_d^t\right) = \sqrt{s_e^2\left(Q_m^v, Q_d^v\right) + s_e^2\left(Q_m^t, Q_d^t\right)} \tag{40}
$$

$$
\begin{aligned}
&s_e\left(Q_m^v \oplus Q_m^t, Q_d^v \oplus Q_d^t\right) = \\
=&\sqrt{s_e^2\left(Q_m^v, Q_d^v\right) - 2s_c\left(Q_m^t, Q_d^t\right) + 2}
\end{aligned}
\tag{41}
$$

$$
\begin{aligned}
&s_e\left(Q_m^v \otimes Q_m^t, Q_d^v \otimes Q_d^t\right) = \\
=&\sqrt{s_e^2\left(Q_m^v, Q_d^v\right)s_c\left(Q_m^t, Q_d^t\right) - 2s_c\left(Q_m^t, Q_d^t\right) + 2}
\end{aligned}
\tag{42}
$$

where the last formula would be a suggested combination choice (Euclidean distance for measuring the similarity between visual representations and cosine similarity for textual representations).

Thus, the standard Rocchio query modification algorithm can be represented as a late fusion, a combination of individual similarity measurements. This late fusion strategy is equivalent to the standard query modification approach.

Table II presents the summary of all the findings. Figures 1 to 11 in the Appendix show examples related to concatenation operator interacting with the inner product, tensor product interacting with the inner product, concatenation operator interacting with the cosine similarity, tensor product interacting with the cosine similarity, weighted concatenation operator interacting with the cosine similarity, concatenation operator interacting with the Euclidean distance, tensor product interacting with the Euclidean distance, concatenation operator interacting with the Bhattacharya similarity, concatenation operator interacting with the Euclidean distance for visual features and cosine similarity for text, tensor product interacting with the Euclidean distance for visual features and cosine similarity for text, and concatenation operator interacting with the Minkowski Family of Distances, respectively.

TABLE II: Summary of the findings.

Early fusion interacting with the similarity	**Late fusion equivalent**
$s_{in}\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right)$	$s_{in}\left(d_1^v \middle| d_2^v\right) + s_{in}\left(d_1^t \middle| d_2^t\right)$
$s_{in}\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right)$	$s_{in}\left(d_1^v \middle| d_2^v\right) \cdot s_{in}\left(d_1^t \middle| d_2^t\right)$
$s_c\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right)$	$s_c\left(d_1^v, d_2^v\right) \cdot s_c\left(d_1^t, d_2^t\right)$
$s_c\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right)$	$\frac{1}{2}\left(s_c\left(d_1^v, d_2^v\right) + s_c\left(d_1^t, d_2^t\right)\right)$
$s_c\left(r_1 d_1^v \oplus r_2 d_1^t, r_1 d_2^v \oplus r_2 d_2^t\right)$	$\frac{\left(r_1^2 s_c\left(d_1^v, d_2^v\right) + r_2^2 s_c\left(d_1^t, d_2^t\right)\right)}{r_1^2 + r_2^2}$
$s_e\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right)$	$\sqrt{s_e^2\left(d_1^v, d_2^v\right) + s_e^2\left(d_1^t, d_2^t\right)}$
$s_e\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right)$	$\sqrt{s_e^2\left(d_1^v, d_2^v\right) + s_e^2\left(d_1^t, d_2^t\right) - \frac{1}{2} s_e^2\left(d_1^v, d_2^v\right) s_e^2\left(d_1^t, d_2^t\right)}$
$s_b\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right)$	$s_b\left(d_1^v, d_2^v\right) + s_b\left(d_1^t, d_2^t\right)$
$s_b\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right)$	$-ln\left(w_1 + w_2\right)$ $w_1 = e^{-s_b\left(d_1^v, d_2^v\right)}$ $w_2 = e^{-s_b\left(d_1^t, d_2^t\right)}$
$s_e\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right)$	$\sqrt{s_e^2\left(d_1^v, d_2^v\right) - 2 s_c\left(d_1^t, d_2^t\right) + 2}$
$s_e\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right)$	$\sqrt{s_e^2\left(d_1^v, d_2^v\right) s_c\left(d_1^t, d_2^t\right) - 2 s_c\left(d_1^t, d_2^t\right) + 2}$
$s_p\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right)$	$\left(s_p^p\left(d_1^v, d_2^v\right) + s_p^p\left(d_1^t, d_2^t\right)\right)^{\frac{1}{p}}$
$tr\left(\left(M_1 \otimes M_2\right) \cdot \left(\left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right)\right)\right)$	$\left(r_1 \cdot \langle q_v \middle| a\rangle^2 + \frac{r_2}{n} \cdot \sum_i \langle c^i \middle| a\rangle^2\right) \cdot \left(r_1 \cdot \langle q_t \middle| b\rangle^2 + \frac{r_2}{n} \cdot \sum_i \langle d^i \middle| b\rangle^2\right)$
$\left\langle P_1 \otimes P_2 \middle| \left(a^T a\right) \otimes \left(b^T b\right)\right\rangle$	$\left\| \left(\left\langle p_1^v \middle| a\right\rangle, \ldots, \left\langle p_n^v \middle| a\right\rangle\right) \otimes \left(\left\langle p_1^t \middle| b\right\rangle, \ldots, \left\langle p_n^t \middle| b\right\rangle\right) \right\|^2$
$\langle Q_m \middle| Q_d\rangle$	$a \langle Q_o \middle| Q_d\rangle + \frac{b}{\lvert D_r\rvert} \sum_{D_j \in D_r} \langle D_j \middle| Q_d\rangle - \frac{c}{\lvert D_{nr}\rvert} \sum_{D_k \in D_{nr}} \langle D_k \middle| Q_d\rangle$
$s_c\left(Q_m, Q_d\right)$	$\frac{1}{\lVert Q_m\rVert}\left(a s_c\left(Q_o, Q_d\right) + \frac{b}{\lvert D_r\rvert} \sum_j s_c\left(D_j, Q_d\right) - \frac{c}{\lvert D_{nr}\rvert} \sum_k s_c\left(D_k, Q_d\right)\right)$
$\lVert Q_m\rVert^2$	$a^2 + c^2 + \frac{2ab}{\lvert D_r\rvert} \sum_j s_c\left(Q_o, D_j\right) - \frac{2ac}{\lvert D_{nr}\rvert} \sum_k s_c\left(Q_o, D_k\right) - \frac{2bc}{\lvert D_r\rvert \cdot \lvert D_{nr}\rvert} \sum_j \sum_k s_c\left(D_j, D_k\right)$
$s_e^2\left(Q_m, Q_d\right)$	$a^2 + c^2 + 2ab - 2ac - 2bc - 2a - 2b - 2c + 1 -$ $\frac{ab}{\lvert D_r\rvert} \sum_j s_e\left(Q_o, D_j\right) + \frac{ac}{\lvert D_{nr}\rvert} \sum_k s_e\left(Q_o, D_k\right) + \frac{bc}{\lvert D_r\rvert \cdot \lvert D_{nr}\rvert} \sum_j \sum_k s_e\left(D_j, D_k\right) +$ $a s_e\left(Q_o, Q_d\right) + \frac{b}{\lvert D_r\rvert} \sum_j s_e\left(D_j, Q_d\right) + \frac{c}{\lvert D_{nr}\rvert} \sum_k s_e\left(D_k, Q_d\right)$

V. Conclusions and Future Work

Fusion strategies are widely utilized in many areas of research, including Information Retrieval. Findings presented in this paper are universal and also apply to other areas of research. Here, however, we focus on the application of fusion strategies to Content-based Image Retrieval (CBIR).

In this paper, we have investigated some interesting interactions between widely used similarity measurements and widely used operators related to early fusion strategy. We have shown that these interactions between specific similarity measurements and specific early fusion strategies have resulted in combinations of representations at the decision level (late fusion strategy). In other words, we have mathematically proved that specific combinations of early fusion strategies and specific similarity measurements are equivalent to particular combinations of measurements (i.e. relevance scores) computed on individual feature spaces.

We have also shown that the query modification method with specific similarity measurements (classic Rocchio algorithm) can be interpreted as weighted combinations of individual similarity measurements. What this mean is that the same effect can be achieved by either modifying the query or combining individual relevance scores. The existing late fusion analogues to the Rocchio algorithm have been considered as separate, different techniques. However, we have seen that the Rocchio model also have its dual late fusion form representations.

For future work we plan to search for other combinations of various operators and similarity measures that could interact in such a way as to represent late fusion. We have discovered that even advanced early fusion can be represented as specific combinations of similarity measurements. We will be also investigating whether the late fusion is capable of capturing the correlation between feature spaces or the interaction between the early fusion operators and the similarity measurements decorrelates features.

Acknowledgment

This work has been partially funded by the EC FP7 OPENi project no. 317883 on Internet of Services (http://www.openi-ict.eu/). The motivation of the project is to provide cloud-based applications and services for mobile users. The AmbieSense work is focused on providing advanced search facilities for large-scale multimedia and social media applications in the cloud.

References

[1] *http://www.courses.ischool.berkeley.edu/i240/s11/Lectures/Lecture_15.ppt* Berkeley - Lecture notes.

[2] P.D. Bruza, K. Kitto, D. Nelson, C.L. McEvoy. Entangling words and meaning. *Proceedings of the 2nd quantum interaction symposium*, 118–124, 2008.

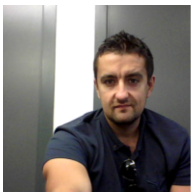
[3] Borth, Damian and Ji, Rongrong and Chen, Tao and Breuel, Thomas and Chang, Shih-Fu. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. *Proceedings of the 21st ACM International Conference on Multimedia*, 223–232, 2013.

[4] N. Bhowmik, V. R. Gonzalez, V. Gouet-Brunet, H. Pedrini, G. Bloch. Efficient fusion of multidimensional descriptors for image retrieval. *IEEE International Conference on Image Processing*, 5766–5770, 2014.

[5] J. C. Caicedo. Multimodal information spaces for content-based image retrieval. *In Proceedings of the Third BCS-IRSG conference on Future Directions in Information Access*, 110–116, 2009.

[6] Y.C. Chang, H.H. Chen. Increasing relevance and diversity in photo retrieval by result fusion. *Working Notes of CLEF 2008*, 2008.

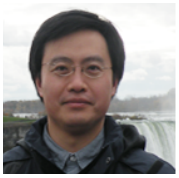
[7] Z. Chen, W. Liu, F. Zhang, M. J. Li and H. J. Zhang. Web mining for web image retrieval. *Journal of the American Society for Information Science and Technology*, 52(10):831–839, 2001.
[8] A. Depeursinge, H. Muller. Fusion techniques for combining textual and visual information retrieval. *ImageCLEF, The Springer International Series on Information Retrieval*, 32:95-114, 2010.
[9] L. Kaliciak, H. Myrhaug, A. Goker, D. Song. On the duality of specific early and late fusion strategies. *Information Fusion (FUSION), 17th International Conference on*, 1–8, 2014.
[10] L. Kaliciak, D. Song, N. Wiratunga, J. Pan. Combining visual and textual systems within the context of user feedback. *The 19th International Conference on Multimedia Modeling*, 7732(1):445-455, 2013.
[11] Kumar, Ashnil and Kim, Jinman and Cai, Weidong and Fulham, Michael and Feng, Dagan. Content-Based Medical Image Retrieval: A Survey of Applications to Multidimensional and Multimodality Data. *Journal of Digital Imaging*, vol. 26, 1025–1039, 2013.
[12] V. Lavrenko, W. B. Croft. Relevance models in information retrieval. *In Croft and Lafferty*, 13:11–56, 2003.
[13] Y.Li, H. Cunningham. Geometric and quantum methods for information retrieval. *SIGIR Forum*, 42(2):22–32, 2008.
[14] F. Lingenfelser, J. Wagner, E. Andre. A systematic discussion of fusion techniques for multi-modal affect recognition tasks. *ICMI*, 19–26, 2011.
[15] H. Liu. A framework for understanding user interaction with content-based image retrieval: model, interface and users. PhD Thesis. *kmi.open.ac.uk/people/alumni/haiming-liu*, 2010.
[16] H. Liu, D. Song, S. Rueger, R. Hu, V. Uren. Comparing dissimilarity measures for content-based image retrieval. *The 4th Asia Information Retrieval Symposium*, 44-50, 2008.
[17] T. Mensink, G. Csurka, F. Perronnin. LEAR and XRCE's participation to visual concept detection task - ImageCLEF 2010. *CLEF 2010 - Conference on Multilingual and Multimodal Information Access evaluation*, 77–80, 2006.
[18] T. Mensink, J. Verbeek, G. Csurkay. Weighted transmedia relevance feedback for image retrieval and auto-annotation. *Technical Report Number 0415*, 2011.
[19] B.A. Olshausen, D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.
[20] A. Popescu-Belis, R. Stiefelhagen. A Probabilistic Model for User Relevance Feedback on Image Retrieval. *Lecture Notes in Computer Science*, 5237:260–271, 2008.
[21] D. Picard, N. Thome, and M. Cord. An efcient system for combining complementary kernels in complex visual categorization tasks. *Proceedings of 17th IEEE International Conference on Image Processing*, 3877-3880, 2010.
[22] E. Rabinovich, O. Rom, O. Kurland. Utilizing relevance feedback in fusion-based retrieval. *In Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval*, 313–322, 2014.
[23] C.J. van Rijsbergen. The geometry of information retrieval. *Cambridge University Press*, 2004.
[24] V. Risojevic and Z. Babic. Fusion of global and local descriptors for remote sensing image classication. *IEEE Geoscience and Remote Sensing Letters*, 10(4):836–840, 2013.
[25] S. Robertson, H. Zaragoza, M. Taylor. Simple BM25 extension to multiple weighted fields. *In Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 42–49, 2004.
[26] J. Rocchio. Relevance feedback in information retrieval. *The Smart Retrieval System-Experiments in Automatic Document Processing*, 313-323, 1971.
[27] Snoek, Cees GM, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. *In Proceedings of the 13th annual ACM international conference on Multimedia*, 399–402, 2005.
[28] Snoek, Cees GM, Marcel Worring, Jan C. Van Gemert, Jan-Mark Geusebroek, and Arnold WM Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. *In Proceedings of the 13th annual ACM international conference on Multimedia*, 421–430, 2006.
[29] Vogt, Christopher C., and Garrison W. Cottrell. Fusion via a linear combination of scores. *Information Retrieval*, 1(3):151–173, 1999.
[30] J. Wang, D. Song, L. Kaliciak. Tensor product of correlated text and visual features: a quantum theory inspired image retrieval framework. *AAAI-Fall 2010 Symposium on Quantum Information for Cognitive, Social, and Semantic Processes*, 109–116, 2010.
[31] D. Widdows. Semantic vector products: Some initial investigations. *Proceedings of the 2nd Quantum Interaction Symposium*, 2008.
[32] *www.quantum.umb.edu/Jacobs/QMT/QMT_AppendixA.pdf* Combining systems: the tensor product and partial trace.
[33] W. Zhang, Z. Qin, and T. Wan. Image scene categorization using Multi-Bag-of-Features. *Proceedings of International Conference on Machine Learning and Cybernetics*, 4:1804–1808, 2011.
[34] Ye, Zheng and He, Ben and Huang, Xiangji and Lin, Hongfei. Revisiting Rocchios Relevance Feedback Algorithm for Probabilistic Models. *Lecture Notes in Computer Science*, 6458:151–161, 2010.
[35] Xie, Shasha, and Yang Liu. Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. *Acoustics, Speech and Signal Processing*, 4985–4988, 2008.



Leszek Kaliciak Dr Leszek Kaliciak works as a Research Scientist within the field of Image and Multimedia Retrieval at AmbieSense Ltd. He researches and develops products and services for image retrieval, mobile applications, and cloud. His research interests include image processing and multimedia retrieval, and information fusion. He is a member of ECIR, SIGIR, AIRS, and ACM Transactions on Multimedia Computing, Communications and Applications. He is leading the image retrieval development for AmbieSense on several projects. For instance, he is currently working on the OPENi EU project where he is researching, developing, and optimizing new and novel search and information fusion technologies for cloud service provision.

Hans Myrhaug He is a specialist in mobile technologies, ubiquitous solutions provision, and co-founder of AmbieSense Ltd. Apart from being the CTO of AmbieSense he also led the application development work of the 14 mill Euro webinos Integrated Project with 70+ people involved across 22 partners. He was also the Coordinator of the AmbieSense EU project, which was a 5.7 mill Euro international project with 50+ people involved throughout the project, with 8 project partners, and several sub-contractors. Its successful completion led to the establishment of AmbieSense, the company, offering software and hardware products. He worked as Project Coordinator and research scientist for the applied research institute SINTEF ICT in Norway, prior to spinning out AmbieSense as a business. He has more than 18 years experience in international research and innovation creating a range of wireless and ubiquitous solutions and services based on wireless sensors, miniaturized devices, search, and the cloud, including more recently also a high-end software camera. As CTO he heads the research and development activities of AmbieSense Ltd across technologies such as software, hardware, and cloud.

Ayse Goker Prof. Goker has over 20 years' research and development experience in areas including context-learning algorithms, multimedia and social media, web user logs, personalization and mobile information systems. Her work involves a strong user-centered approach to algorithm and search-system design, development and evaluation. Goker has been involved in the webinos and the Social Sensor Integrated Projects. Social Sensor is enabling real-time multimedia indexing and search in the Social Web, with partners such as Yahoo, Alcatel-Lucent, Deutsche Welle, IBM, and ITI-Certh. She has been an entrepreneur and pioneer in context-aware information retrieval systems since the early 1990s. She holds a lifelong Enterprise Fellowship from the Royal Society of Edinburgh and Scottish Enterprise for her contributions to the AmbieSense EU project, which subsequently commercialized as AmbieSense Ltd. She is also Professor of Computational Systems, School of Computing, Robert Gordon University, chairing the Digital Technologies Theme at IDEAS research center.



Dawei Song Prof. Dawei Song obtained his PhD from the Chinese University of Hong Kong in year 2000. He worked as a Research Scientist at the Distributed Systems Technology Center, Australia, and as a Professor of Computing at the Robert Gordon University in Scotland. He currently works at the Open University in Milton Keynes, England, and the Tianjin University in Tianjin, China. His major research interests include information retrieval models, web and enterprise search systems, biomedical information retrieval, text-based knowledge discovery from biomedical resources, and management and search of complex data (multimedia and biological structures). He has been a Principal Investigator of a number of prestigious research projects funded by the UK's Engineering and Physical Sciences Research Council, and has received twice the IBM Innovation Awards in 2006 and 2007. He is a professional member of ACM, and the Secretary of the Information Retrieval Specialists Group of the British Computer Society (BCS). He is a Steering Committee Member of the Asian Information Retrieval Symposium (AIRS). He is on editorial board of the International Journal of Computer Processing of Languages. He has served as a guest editor for the ACM Transactions on Asian Language Information Processing, and as a Programme Committee member for a number of well-respected conferences such as ACM SIGIR, CIKM, and ECIR. He has published papers on respected journals such as ACM transactions on Information Systems, Journal of American Society for Information Science and Technology, Decision Support Systems, IEEE Transactions on Knowledge and Data Engineering, and top conferences such as ACM SIGIR and CIKM.

VI. Appendix

Example 1. Concatenation with the inner product.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = \left\langle \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \oplus \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right) \mid (1,0) \oplus (1,0,0) \right\rangle =$
$\left\langle \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2} | 1, 0, 1, 0, 0 \right\rangle = \sqrt{2}$
$R = \left\langle \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \mid (1,0) \right\rangle + \left\langle \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right) \mid (1,0,0) \right\rangle =$
$\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2} = \sqrt{2}$
$L = R$
Therefore the concatenation with the inner product as a similarity measurement can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 2. Tensor product with the inner product.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = \left\langle \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \otimes \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right) \mid (1,0) \otimes (1,0,0) \right\rangle =$
$\left\langle \frac{1}{2}, 0, \frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2} | 1, 0, 0, 0, 0, 0 \right\rangle = \frac{1}{2}$
$R = \left\langle \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \mid (1,0) \right\rangle \cdot \left\langle \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right) \mid (1,0,0) \right\rangle =$
$\frac{\sqrt{2}}{2} \cdot \frac{\sqrt{2}}{2} = \frac{1}{2}$
$L = R$
Therefore the tensor product with the inner product as a similarity measurement can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 3. Concatenation with the cosine similarity.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = s_c\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \oplus \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1,0) \oplus (1,0,0)\right) =$
$\frac{\sqrt{2}}{\sqrt{\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}} \cdot \sqrt{1+1}} = \frac{\sqrt{2}}{2}$
$R = \frac{1}{2} \cdot \left(s_c\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), (1,0)\right) + s_c\left(\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1,0,0)\right)\right) =$
$\frac{1}{2} \cdot \left(\frac{\frac{\sqrt{2}}{2}}{1 \cdot 1} + \frac{\frac{\sqrt{2}}{2}}{1 \cdot 1}\right) = \frac{\sqrt{2}}{2}$
$L = R$
Therefore the concatenation with the cosine similarity as a similarity measurement can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 4. Tensor product with the cosine similarity.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = s_c\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \otimes \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0) \otimes (1, 0, 0)\right) =$
$s_c\left(\left(\frac{1}{2}, 0, \frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2}\right), (1, 0, 0, 0, 0, 0)\right) = \frac{\frac{1}{2}}{1 \cdot 1} = \frac{1}{2}$
$R = s_c\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), (1, 0)\right) \cdot s_c\left(\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 0)\right) =$
$\frac{\frac{\sqrt{2}}{2}}{\sqrt{\frac{1}{2}+\frac{1}{2}} \cdot \sqrt{1}} \cdot \frac{\frac{\sqrt{2}}{2}}{\sqrt{\frac{1}{2}+\frac{1}{2}} \cdot \sqrt{1}} = \frac{\sqrt{2}}{2} \cdot \frac{\sqrt{2}}{2} = \frac{1}{2}$
$L = R$
Therefore the tensor product with the cosine similarity as a similarity measurement can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 5. Weighted concatenation with the cosine similarity.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = s_c\left(2 \cdot \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \oplus 4 \cdot \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), 2 \cdot (1, 0) \oplus 4 \cdot (1, 0, 0)\right) =$
$s_c\left(\left(\sqrt{2}, \sqrt{2}, 2\sqrt{2}, 0, 2\sqrt{2}\right), (2, 0, 4, 0, 0)\right) =$
$\frac{2\sqrt{2}+8\sqrt{2}}{\sqrt{2+2+8+8} \cdot \sqrt{4+16}} = \frac{10\sqrt{2}}{\sqrt{20} \cdot \sqrt{20}} = \frac{\sqrt{2}}{2}$
$R =$
$\frac{1}{4+16} \cdot \left(4s_c\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), (1, 0)\right) + 16s_c\left(\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 0)\right)\right) =$
$\frac{1}{20} \cdot \left(4\frac{\sqrt{2}}{2} + 16\frac{\sqrt{2}}{2}\right) = \frac{\sqrt{2}}{2}$
$L = R$
Therefore the weighted concatenation with the cosine as a similarity measurement can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 6. Concatenation with the Euclidean distance.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = s_e\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \oplus \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0) \oplus (1, 0, 0)\right) =$
$s_e\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 1, 0, 0)\right) =$
$\sqrt{\left(\frac{\sqrt{2}}{2} - 1\right)^2 + \frac{1}{2} + \left(\frac{\sqrt{2}}{2} - 1\right)^2 + \frac{1}{2}} = \sqrt{4 - 2\sqrt{2}}$
$R = \sqrt{s_e^2\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), (1, 0)\right) + s_e^2\left(\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 0)\right)} =$
$\sqrt{\left(\frac{\sqrt{2}}{2} - 1\right)^2 + \left(\frac{\sqrt{2}}{2}\right)^2 + \left(\frac{\sqrt{2}}{2} - 1\right)^2 + \left(\frac{\sqrt{2}}{2}\right)^2} = \sqrt{4 - 2\sqrt{2}}$
$L = R$
Therefore the concatenation with the Euclidean distance as a similarity measurement can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 7. Tensor product with the Euclidean distance.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = s_e\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \otimes \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0) \otimes (1, 0, 0)\right) =$
$s_e\left(\left(\frac{1}{2}, 0, \frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2}\right), (1, 0, 0, 0, 0, 0)\right) = \sqrt{\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}} = 1$
$R^2 = s_e^2\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), (1, 0)\right) + s_e^2\left(\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 0)\right) -$
$\frac{1}{2}s_e^2\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), (1, 0)\right) \cdot s_e^2\left(\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 0)\right) =$
$2 - \sqrt{2} + 2 - \sqrt{2} - \frac{1}{2}(2 - \sqrt{2})^2 = 1$
$R = \sqrt{1} = 1$
$L = R$
Therefore the tensor product with the Euclidean distance as a similarity measurement can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 8. Concatenation with the Bhattacharya similarity.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = s_b\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \oplus \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0) \oplus (1, 0, 0)\right) =$
$s_b\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 1, 0, 0)\right) =$
$-ln\left(\sqrt{\frac{\sqrt{2}}{2}} + \sqrt{\frac{\sqrt{2}}{2}}\right) = -ln\left(2 \cdot 2^{-\frac{1}{4}}\right) = -ln\left(2^{\frac{3}{4}}\right)$
$R = s_b\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), (1, 0)\right) + s_b\left(\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 0)\right) =$
$-ln\left(e^{ln\left(\sqrt{\frac{\sqrt{2}}{2}}\right)} + e^{ln\left(\sqrt{\frac{\sqrt{2}}{2}}\right)}\right) = -ln\left(2\sqrt{\frac{\sqrt{2}}{2}}\right) = -ln\left(2^{\frac{3}{4}}\right)$
$L = R$
Therefore the concatenation with the Bhattacharya similarity as a similarity measurement can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 9. Concatenation with the Euclidean distance for visual features and cosine similarity for text.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = s_e\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \oplus \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0) \oplus (1, 0, 0)\right) =$
$s_e\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 1, 0, 0)\right) =$
$\sqrt{4 - 2\sqrt{2}}$
$s_e\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), (1, 0)\right) = \sqrt{2 - \sqrt{2}}$
$s_c\left(\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 0)\right) = \frac{\sqrt{2}}{2}$
$R = \sqrt{2 - \sqrt{2} - \sqrt{2} + 2} = \sqrt{4 - 2\sqrt{2}}$
$L = R$
Therefore the concatenation with the above similarity measurements can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 10. Tensor product with the Euclidean distance for visual features and cosine similarity for text.

Query visual representation: $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
Query textual representation: $\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$
Arbitrary image from the collection (visual): $(1, 0)$
Arbitrary image from the collection (text): $(1, 0, 0)$
$L = s_e\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \otimes \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0) \otimes (1, 0, 0)\right) =$
$s_e\left(\left(\frac{1}{2}, 0, \frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2}\right), (1, 0, 0, 0, 0, 0)\right) = \sqrt{\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}} = 1$
$s_e\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right), (1, 0)\right) = \sqrt{2 - \sqrt{2}}$
$s_c\left(\left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right), (1, 0, 0)\right) = \frac{\sqrt{2}}{2}$
$R = \sqrt{\left(2 - \sqrt{2}\right) \cdot \frac{\sqrt{2}}{2} - \sqrt{2} + 2} = 1$
$L = R$
Therefore the tensor product with the above similarity measurements can be represented in a late fusion form.
This means that these specific early and late fusion strategies must produce the same ranking of images.

Example 11. Concatenation with the Minkowski Family of Distances.

Query visual representation: $d_1^v = (1, 3, 4)$
Query textual representation: $d_1^t = (12, 1, 4, 2)$
Arbitrary image from the collection (visual): $d_2^v = (0, 3, 5)$
Arbitrary image from the collection (text): $d_2^t = (11, 0, 3, 1)$
$s_{p=\frac{1}{4}}\left(d_1^v, d_2^v\right) =$
$\left(|1 - 0|^{\frac{1}{4}} + |3 - 3|^{\frac{1}{4}} + |4 - 5|^{\frac{1}{4}}\right)^4 = 2^4 = 16$
$s_{p=\frac{1}{4}}\left(d_1^t, d_2^t\right) =$
$\left(|12 - 11|^{\frac{1}{4}} + |1 - 0|^{\frac{1}{4}} + |4 - 3|^{\frac{1}{4}} + |2 - 1|^{\frac{1}{4}}\right)^4 = 4^4 = 256$
Therefore, the right-hand side of the equation becomes
$R = \left(s^{\frac{1}{4}}_{p=\frac{1}{4}}\left(d_1^v, d_2^v\right) + s^{\frac{1}{4}}_{p=\frac{1}{4}}\left(d_1^t, d_2^t\right)\right)^4 =$
$\left(16^{\frac{1}{4}} + 256^{\frac{1}{4}}\right)^4 = (2 + 4)^4 = 1296$
For the left-hand side, we have
$L = s_{p=\frac{1}{4}}\left(d_1^v \oplus d_1^t, d_2^v \oplus d_2^t\right) =$
$s_{p=\frac{1}{4}}\left((1, 3, 4, 12, 1, 4, 2), (0, 3, 5, 11, 0, 3, 1)\right) =$
$(1 + 0 + 1 + 1 + 1 + 1 + 1)^4 = 1296$
Thus, $L = R$.
Therefore concatenation operation with the above similarity measurements can be represented in a late fusion form.
This means that this specific early and late fusion strategy must produce the same ranking of images.